IS733 Data Mining BRFss - Diabetes

Sushmitha Kanapuram
David Calderone
Sanju Biju
Maulik Soni
Sudhanshu Jha
Rohith Boggarapu

Introduction & Objective

The purpose and goal of this research was to find a large dataset of the United States (US) population and analyze it using the Waikato Environment for Knowledge Analysis (Weka) to validate known causal and correlational factors. A secondary purpose was to see if any new factors emerged.

The motivation of our finding is by 2030 to 2045 estimated number of diabetes worldwide could reach 550 million to 800 million in in number. .

Diabetes is believed to have first been noted by the Egyptians in the mid-1500s (Porter, 2018). Those inflicted were observed to have urine termed “honey urine” as bees were attracted to the high sugar content of the urine, a physiological outcome of high blood sugar where sugar essentially “spills” out of the kidneys into the urine.

Diabetes can be classified into two types – Type 1 Diabetes and Type 2 Diabetes (T2D). Diabetes is a long-term (chronic) illness that affects how your body converts food into energy. The majority of the food you consume is converted by your body into sugar (glucose), which is then released into your bloodstream. Your pancreas releases insulin when your blood sugar levels rise. In order for blood sugar to enter your body's cells and be used as energy, insulin functions like a key.

BRFss

Two biggest surveys conducted in USA for collecting the health data is BRFSS (Behavioral risk factor surveillance system) and NHIT (National Health interview surveys)

The primary reason for selecting BRFSS over NHIS.

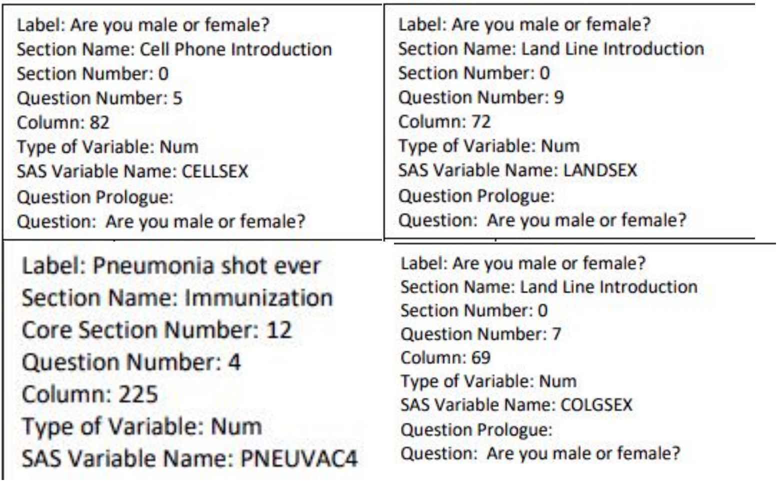
- the BRFSS provides national estimates comparable to those of the NHIS. BRFSS national data could provide rapidly available information to guide national policy and program decisions.

- The questionnaire updated every 15-20 years for NHIS, whereas for BRFSS as it’s telephone survey system, it updates every year.

- NHIS conduct the data based on household, so the persons don’t have fixed household address excluded from the survey which creates biased and implicit incomplete survey.

Data Preprocessing

- 401,958 instances, 279 attributes - potentially 112M values
- CDC - Codebook; a Data Dictionary
- XPT File -> CSV through Python



Label: Are you male or female?
Section Name: Cell Phone Introduction
Section Number: 0
Question Number: 5
Column: 82
Type of Variable: Num
SAS Variable Name: CELLSEX
Question Prologue:
Question: Are you male or female?

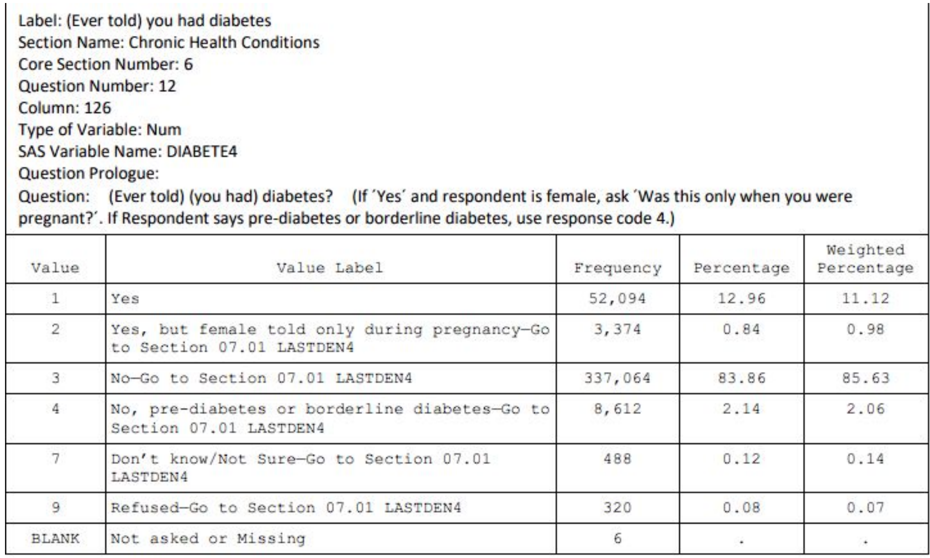
Label: Are you male or female?
Section Name: Land Line Introduction
Section Number: 0
Question Number: 9
Column: 72
Type of Variable: Num
SAS Variable Name: LANDSEX
Question Prologue:
Question: Are you male or female?

Label: Pneumonia shot ever
Section Name: Immunization
Core Section Number: 12
Question Number: 4
Column: 225
Type of Variable: Num
SAS Variable Name: PNEUVAC4

Label: Are you male or female?
Section Name: Land Line Introduction
Section Number: 0
Question Number: 7
Column: 69
Type of Variable: Num
SAS Variable Name: COLGSEX
Question Prologue:
Question: Are you male or female?

- Cleaning
  - Blank / Missing / Refused to answer instances
- No integration
  - sole source data
- Transformation
  - CDC Codebook provides value tables for each attribute's true meaning
- Reduction
  - Redundant attributes questions - in different formats with different results
  - Irrelevant attributes
- Results in a narrowed data set, of 12 Attributes
  - Reducing instances further

Class Attribute:

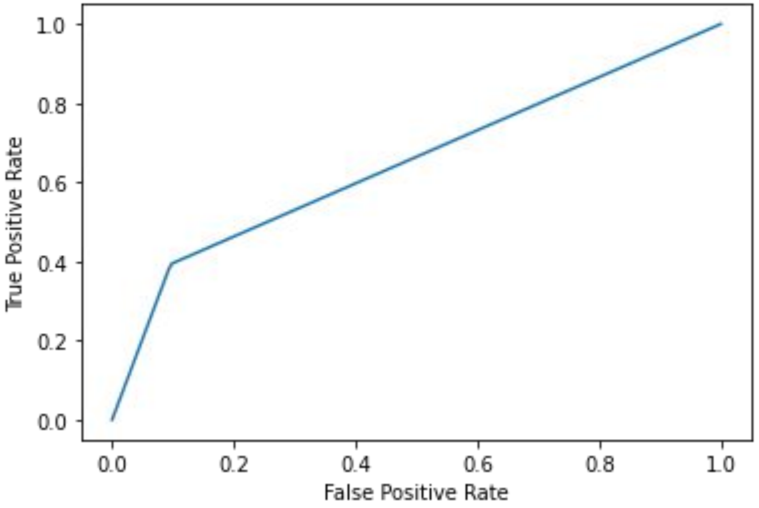


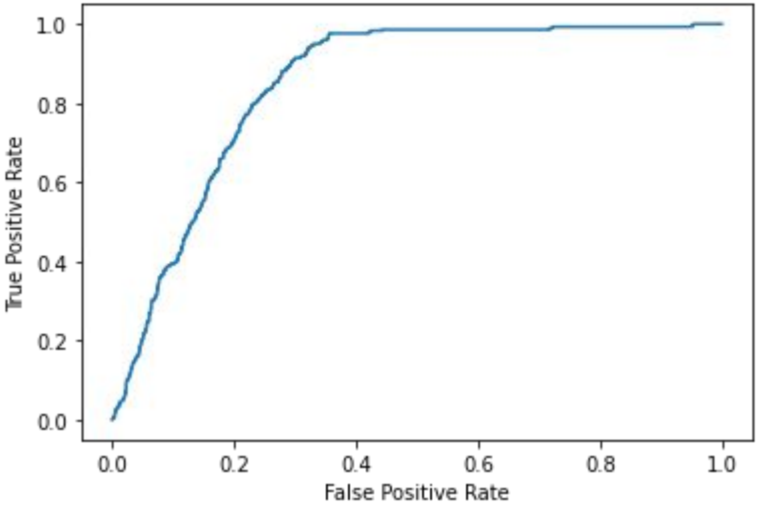
Label: (Ever told) you had diabetes
Section Name: Chronic Health Conditions
Core Section Number: 6
Question Number: 12
Column: 126
Type of Variable: Num
SAS Variable Name: DIABETE4
Question Prologue:
Question: (Ever told) (you had) diabetes? (If 'Yes' and respondent is female, ask 'Was this only when you were pregnant?'. If Respondent says pre-diabetes or borderline diabetes, use response code 4.)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	52,094	12.96	11.12
2	Yes, but female told only during pregnancy—Go to Section 07.01 LASTDEN4	3,374	0.84	0.98
3	No—Go to Section 07.01 LASTDEN4	337,064	83.86	85.63
4	No, pre-diabetes or borderline diabetes—Go to Section 07.01 LASTDEN4	8,612	2.14	2.06
7	Don't know/Not Sure—Go to Section 07.01 LASTDEN4	488	0.12	0.14
9	Refused—Go to Section 07.01 LASTDEN4	320	0.08	0.07
BLANK	Not asked or Missing	6	.	.

Analytics

-> After preliminary data cleaning we did some data preprocessing by changing data types from float or int to categorical data wherever need using python.
-> Due to computing limitation we took random 10,000 data and run three supervised learning model Logistic Regression , Decision tree and Gaussian Naive Bayes.
-> Here are the roc curve for two most efficient model .

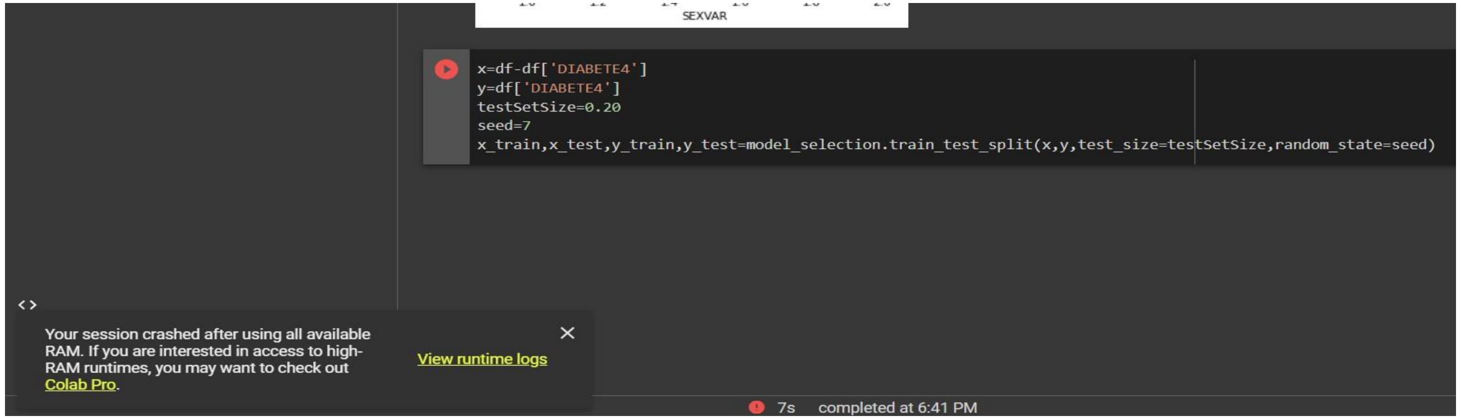




Challenges

We faced a few challenges while performing in order to complete the project.

Because of the humongous data that we choose for the project, data cleaning part was a major task to complete.
As we have chose the BRFSS data it did have the other contents like frequency of alcohol and smoking which is irrelevant.
Refer to the below screenshot.



```
x=df-df['DIABETE4']
y=df['DIABETE4']
testSetSize=0.20
seed=7
x_train,x_test,y_train,y_test=model_selection.train_test_split(x,y,test_size=testSetSize,random_state=seed)
```

Your session crashed after using all available RAM. If you are interested in access to high-RAM runtimes, you may want to check out Colab Pro.

7s completed at 6:41 PM

Limitation and future work

- Limitation of our result will be we are not differentiating between type 1 and type 2 diabetes. So I will be ambiguous to select the precise predictors/attributes causing diabetes.

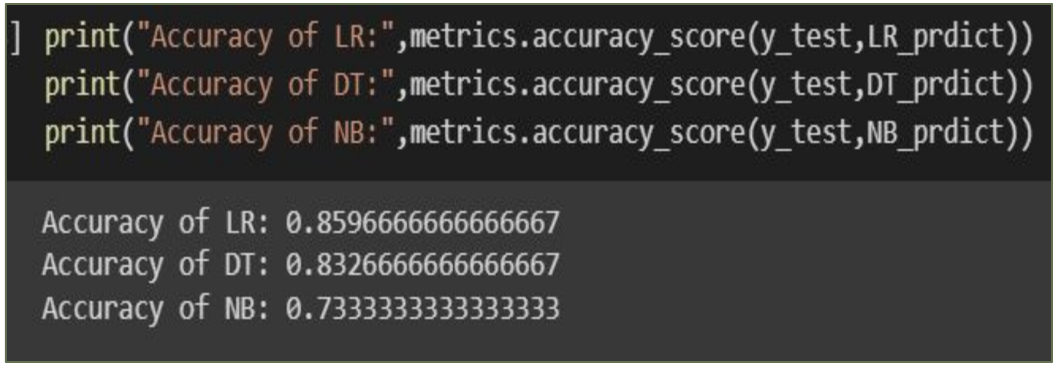
- The previous year is excluded from the dataset so it just a description of small snapshot of whole population over the years, and won't include the aggregation of health related issue over the years.

- In future, It will be a valuable contribution to find the diabetes predictors based on the zip code. Since it is possible to include the specific population over telephonic survey, we can can get the area wise prediction.

- Lifestyles also holds significant part for causing the diabetes, research based on the lifestyle behaviours would yield powerful study. E.g. Regular exercise, walking, get regular checkups, etc.

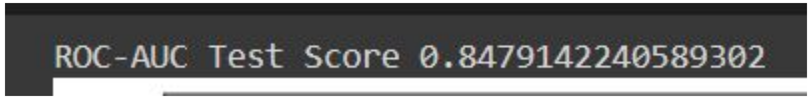
Conclusion

- We performed Logistic Regression, Decision Tree Classifier, and Gaussian naive Bayes Algorithm.
- Among them, logistic regression provided accuracy of 85%.
- We got ROC-AUC score 0.84



```
] print("Accuracy of LR:",metrics.accuracy_score(y_test,LR_prdict))
  print("Accuracy of DT:",metrics.accuracy_score(y_test,DT_prdict))
  print("Accuracy of NB:",metrics.accuracy_score(y_test,NB_prdict))
```

```
Accuracy of LR: 0.8596666666666667
Accuracy of DT: 0.8326666666666667
Accuracy of NB: 0.7333333333333333
```



```
ROC-AUC Test Score 0.8479142240589302
```